International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 7, Issue 2, Apr 2017, 17-24

AN APPROACH TO EMAIL CATEGORIZATION FOR TELECOMMUNICATION CORPUS

RAJWANT KAUR[1] & GAURAV PATHAK[2]

[1]*Research Scholar, CSE, Chandigarh University, Ajitgarh, India*

[2]*Assistant Professor, Department of Applied Science, Chandigarh University, Ajitgarh, India*

ABSTRACT

At present, most of the transactions and business is taking place through emails and now it is also necessary for log-in any site. Due to this a large number of emails are collected in our email account which is hard to read, manage. That is reason for email categorizing. Classifying those emails into categories is a convenient way for people to read them. So the main aim of this paper is to solve the problem of email overloading by automatically classifies the e-mail into different classes based on the content of e-mail. Telecommunication industry dataset is used for the categorization. So this system classifies the email system into two categories service and finance.

KEYWORDS: Email Mining & Classification

Received: Jan 17, 2017**; Accepted:** Feb 28, 2017; **Published:** Mar 04, 2017**; Paper Id.:** IJCSEITRAPR20173

Original Article

1. INTRODUCTION

With the expansion of networks, emails have become effectual, speedy and most prudent forms of communication. Electronic mail is the method for sending digital messages from one person to another between computers via a network. [1] Email remnant the most ubiquitous form of communication because of moderate cost and massive use of the internet. Emails are pre-eminent for signing any social media site, for shopping online, for online transaction and for online communication. So the number of email users is continually intensified. Acc. to radicati group's report, there are currently 2.6 billion [2] email active users and by the end of 2019 its growth will hike up to over 2.9 billion. But the widening of email accounts is growing slightly faster than the number of email users for the reason that the users have multiple accounts. The proportion of widening of email accounts is 7% per year. But with the growth of sending email messaging, there has been also substantial growth in unsolicited mail. The average number of graymail received per user is fourteen which would exceed to nineteen by the end of 2019. So there is a requisite of email mining. Email mining is not obligatory for graymail filtrating. Despite that, it is also imperative for email foldering. Today we send and receive 90 messages per day. For some people, it is usually more than hundred. Hence users spend a lot of their working time on processing the emails and organize these emails. At the same time, a large part of email traffic consists of business emails, non-personal emails, and friend's emails. People tussle to distinct crucial messages that urging instant attention. So overloading can be tackled by two ways – by email summarization and other is by automatic categorization. Therefore it is uncomplicated to find and organize both incoming and existing emails.

So in this paper Section 2 reviews the previous work in email mining. Section 3 explains the algorithms, techniques and dataset that are used in the previous paper in the tabular form. Section 4 presents the results and

section 5 concludes the paper.

2. LITERATURE SURVEY

Two fast machine learning algorithms [3] that are TF-IDF and Naïve Bayes [NB] are implemented for the email categorization and three categories are made. Both the algorithms are contrasted. NB gives good results than the TF-IDF.

Klimt et al. [4] presented the work on email classification based on relationship data. Experiment is conducted on enron corpus using the SVM classification algorithm. Hence to bring out the terms of emails, parsing is applied and after that using the ltc formula, weights is assigned. Assessment is done on the premise of F1. CMU dataset is put in which is self-created by the author to check the performance of the enron. Results are almost similar.

This technical report [5] presents the email categorization based on the timeline using different supervised learning algorithms. Two large corpuses that are enron and SRI are used for this task. So in the preliminary processing step, the folders which have fewer messages are deleted. Wide margin Winnow algorithm takes less running time as comparative to other algorithms. It is also noticed that wide margin algorithm also outperforms when it is compared to regular winnow.

Xia et al. [6] categorized the emails into the 15 folders for the trouble free access. So for this task two tournament methods are proposed, namely Round Robin Tournament (RRT) and Elimination Tournament (ET). Firstly both the tournament methods are contrasted with n-way classification method, in which tournaments methods gives the higher accuracy. After that ET and RRT are compared in which RRT performs slightly better than ET.

[7]In this paper different classification algorithms are compared which includes J48 decision tree, NB, NN and SVM for the spam mail filtering.

Li and his team introduce [8] ME model and follows two phase way to categorize the emails bases on the contents and properties. Then Li started with preprocessing the mails by filtering the non-character symbol, by resolving the links. In two phase method first it classify the mails into legitimate and spam and in the second phase emails are categorized into 7 categories. For the comparison ME model is tested with NB, SVM, and KNN. ME model is the best one.

This paper [9] implemented the Evolving Email Clustering Method [EECM] that groups the emails based on user's activities. To examine the grouping accuracy of EECM algorithm Davis Bouldin validity index is used, which are used for measuring the goodness, quality, validity of the grouping technique. So EECM algorithm is compared with K-means, Fuzzy over the Enron dataset, in which EECM performs better.

Lu et al. [10] proposes the Semantic Vector Space Model [sVSM] for the purpose of email categorization. The traditional VSM do not contain the semantic relations, so that why the author proposes sVSM method to remove this problem. So for creating semantic vector, features are extracted by considering the hepernymy-hyponymy relations between the synonym sets. To assign the weights of sematic vector tf*iwf*iwf algorithm is used. Three experiments are designed to evaluate the performance of this method. In the first experiment the traditional VSM and sVSM are compared, in which proposed method performed well. In the second experiment, proposed method is contrast with Bayesian and KNN algorithms, in which KNN gives higher results as contrast to others. Third experiment shows that with increasing of email set, categorizing performance also increases.

Matwin [11] presented the co-training algorithm to solve the problem of unlabeled data by using 1500 emails on SVM and Naïve Bayes. Experiment is conducted using the weka tool. Firstly pre-processing is done by removing the stop words and by stemming. Problem is divided into highly balanced, medium balanced and balanced. SVM gives better results than Naïve Bayes. Author also tries to find the reason that why Naïve Bayes gives poor results than SVM. So they experiment with the Naïve Bayes by removing the features. We see that SVM also outperforms with very large features than Naïve Bayes.

Yang et al. [12] discusses about the spam mails in the healthcare organization. So to classify the emails into spam or ham, common characteristics are extracted like no drug effects, disease name from the Trec dataset. Different machine learning algorithms are applied. So to improve the accuracy, Decision tree and naïve Bayes algorithms are combined and it gives higher accuracy as compared to others and also error rate is low.

Kumar et al. [13] compares the fifteen classification algorithms for the classification of email spam. [14] Soni et al. proposes an AEMS (Automatic Email Management System) for the handling of emails.

Mishra et al. [15] also worked on spam categorization. So, to carry out this task author uses the different tools to find out the best one. Weka performs better as compares to Rapid miner and support vector machine.

Tang et al. [16] presents the survey on email mining. Author not only reviewed the single task in the email mining, rather he presented the five major tasks -namely spam detection, contact analysis, email filing, email visualization and email network property analysis. He also mentions the related techniques and software tools to mine the email. Future directions are provided by giving the two examples that are email egocentric network and email monetization.

Many classification algorithms are used for the classification of emails to check whether it is legitimate or non-legitimate. Author [17] performs this experiment in the real environments to check the performance of these algorithms. So the author collected the email datasets from the university, company, research institute. Results show that university gives the higher percentage of spam messages due to various subscription services. Decision tree and SVM gives the better results.

Alsmadi et al. [18] carried out email categorization on the personal email dataset. SVM, KNN, N-gram methods are developed to achieve clustering and classification of emails. Classification based on N-Gram is shown to be the best as text is Bi-language.

Table 1: Various Techniques, Algorithms for Email Mining from the Era 2002 to 2015

Paper	Year	Dataset	Techniques	Algorithms
[3]	2002	R_T	C_f	TF-IDF, NB
[4]	2004	E_r, CMU	C_f	SVM
[5]	2004	E_r, SRI	C_f	NB,SVM, ME, WMW
[6]	2005	R_T	T_m	E_m, RRT
[7]	2007	R_T	C_f	NN, SVM, NB, J48
[8]	2007	R_T	C_f	ME, KNN, SVM, WMW, NB
[9]	2009	R_T	C_r	EECM, K-means
[10]	2010	20-ng	C_f	tf*iwf*iwf, sVSM, KNN, NB
[11]	2011	R_T	C_f	Co_T, SVM, NB
[12]	2012	TC	A_s, C_f, C_r	NB, SVM, J48, K_m
[13]	2012	S_b	C_f	ID3, K-NN, SVM, RF, NB, LDA
[14]	2013	20-ng	A_s, C_r, C_f	A_p, non-parametric K_m++
[15]	2014	U_n, E_r, SA	C_f	RF, B_g, SVM, NB

Table 1: Contd.,				
[17]	2015	R_T	C_f	SVM, J48, NB
[18]	2015	R_T	C_f, C_r	SVM, K_m, NG

R_T-Real time, E_r-Enron, ng-newsgoup, TC-Trec Corpus, S_b-Spambase, U_n-Usenet, SA-Spam Asian, C_f-Classification, C_r-Clustering, A_s-Association, T_m-Tournament, TF-IDF-Term Frequency –Inverse Document Frequency, NB-Naïve Bayes, SVM-Support Vector Machine, NN-Neural Network, WMW-Wide Margin Winnow, KNN-k-Nearest Neighbor, ME-Maximum Entropy, E_m-Elimination, RRT-Round Robin tournament, A_p-Apriori, EECM- Evolving Email Clustering Method, K_m-K-mean, Co_T-Cotraining algorithm RF- Random Forest, B_g-Bagging, NG-NGrams

4. EXPERIMENT

- **Corpus**

In this paper, a telephonic industry's emails are collected. Common data about the emails' dataset is composed from Google provided for categorize the emails based on their content. There are many other public email corpuses available like enron, spambase, usenet, SRI etc. But some corpuses are used to classify the spam emails and some are categorized based on the users. So here we build our own dataset.

- **Emails Content Pre-Processing**

A MIME parser is then used to parse information from those emails to make a dataset that contain one record for every email with the following information parsed: Email file name, email body, subject

- **Emails Content Data Mining**

An automated tool is to further analyze the content from all emails and measure frequency of words. More than 20,000 words are collected. Stemming is also applied in the term frequency table.

- **E-Mail Clustering**

We obtain entire email as centroid and divide into clusters. After dividing into clusters the content, it will pass through the knowledge dictionary set for scanning. We obtain the score for each cluster. Finally calculate the distance between cluster to cluster and cluster to original content.

RESULTS

Table 2

Words	Cluser1	Cluster2	Email
refund	0.67	0.76	1.43
connection	0.89	0.9	1.79
waiver	1.1	0.23	1.33
billing	0.65	1	1.65
charge	0	0.35	0.35
network	0.78	0.9	1.68
prepaid	1.1	0.9	2
postpaid	0.98	0.99	1.97
service	0.87	0.92	1.79
complaint	0.87	0.76	1.63
issue	0.34	0.54	0.88

Figure 1: Shows the IDF Values

Table 3: Shows the Accuracy, Precision, and Recall

PROPERTY	RESULTS
True Positive	955
True Negative	10
False Positive	15
False Negative	0
Sensitivity (Recall)	97.94%
Precision (Positive Predictive Value)	98.45%
Result Prevalence	97.50%
Accuracy	96.50%

For the assessment different metrics Precision, Recall, Accuracy are used.

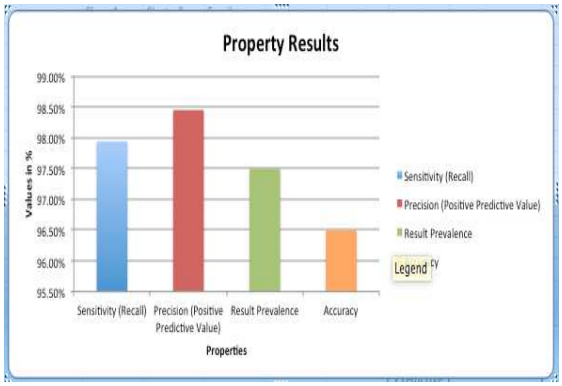

Figure 2: Graphical Representation of Precision, Recall

Work Flow Diagram

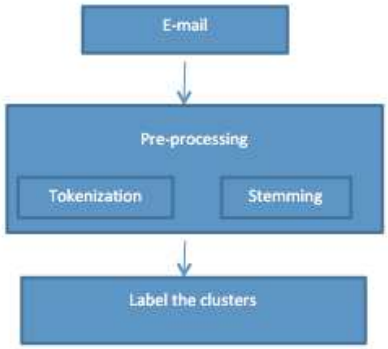

Figure 3: Shows E-Mail Pre-Processing

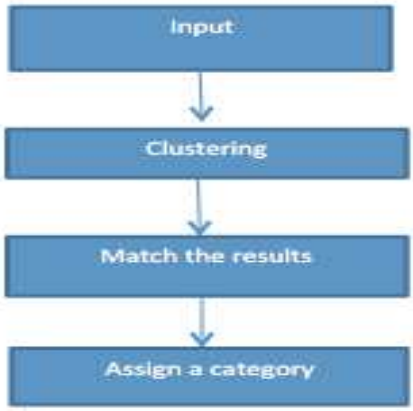

Figure 4: Content Clustering

CONCLUSIONS

Emails' classification in particular utilizes several data mining activities such as: Text parsing, stemming, classification, clustering. There are many goals or reasons why to cluster or classify emails, his may include reasons such as: Spam detection, contact analyses, email categorization. Results show that our system works perfectly by categorizing the email into relevant folders.

REFERENCES

1. http://www.dictionary.com
2. www.radicati.com
3. *Yang, Jihoon and Sung-Yong Park, "Email categorization using fast machine learning algorithms", Print ISBN: 978-3-540-00188-1, pp. 316-323, Springer Berlin Heidelberg, 2002*
4. *Klimt, Bryan and Yiming Yang," The Enron Corpus: A new Dataset for Email Classification Research, " In Machine learning: ECML, Print ISBN: 978-3-540-23105-9, pp.217-226, Springer Berlin Heidelberg, 2004*
5. *Bekkerman, Ron "Automatic categorization of email into folders: Benchmark experiment on Enron and SRI corpora", 2004*
6. *Xia, Yunqing, Wei Liu, and Louise Guthrie. "Email categorization with tournament methods." In Natural Language Processing and Information Systems, Print ISBN: 978-3-540-26031-8, pp. 150-160. Springer Berlin Heidelberg, 2005*
7. *Youn, Seongwook and Dennis McLeod. "Comparative study for Email Classification", Advances and Innovations in Systems, Computing Sciences and Software Engineering, Print ISBN: 978-1-4020-6263-6, pp.387-391, Springer Netherlands, 2007.*
8. *Li, Peifeng, Jinhui Li, and Qiaoming Zhu., "An approach to Email Categorization with the ME Model", in FLAIRS Conference, pp. 229-234, 2007*
9. *Ayodele, Taiwo, Shikun Zhou and Rinat Khausainov," Evolving email clustering method for email grouping: A machine learning approach", In Applications of Digital Information and Web Technologies, ICADIWT'09. Second International Conference on the, E-ISBN: 978-1-4244-4457-1 pp.357-362, IEEE, 2009*
10. *Lu, Zhao and Jianguo Ding, "An efficient semantic VSM based email categorization method", International Conference on Computer Application and System Modeling (ICCASM 2010),Vol.11, E-ISBN: 978-1-4244-7273-6, ISSN :2161-9069, pp. 525-530, IEEE, 2010*

11. *Kiritchenko, Svetlana and Stan Matwin, "Email classification with co-training", In Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research, pp. 301-312. IBM Corp., 2011.*
12. *Yang, Weiwen and Linchi Kwok, "Comparison Study of Email Classification for Health Organizations", International Conference on Information Management, Innovation Management and Industrial Engineering,Vol.3, ISSN: 2155-1456 pp.468-473, IEEE, 2012*
13. *Kumar, R. Kishore, G. Poonkuzhali, and P. Sudhakar. "Comparative study on email spam classifier using data mining techniques." In the proceedings of the International Multi Conference of Engineers and Computer Scientists, vol. 1, pp. 14-16, 2012*
14. *Soni, Gunjan, and C. I. Ezeife. "An automatic Email Management Approach Using Data Mining Techniques." In Data Warehousing and Knowledge Discovery, print ISBN: 978-3-642-40130-5, Series Volume: 8057, online ISBN: 978-3-642-40131-2, pp. 260-267, Springer Berlin Heidelberg, 2013.*
15. *Mishra, Ravishankar, and Ramjeevan Singh Thakur, "An efficient approach for Supervised Learning Algorithms using different data mining tools for spam categorization", In Communication Systems and Network Technologies (CSNT), pp.472-477, IEEE, 2014.*
16. *Tang, Guanting, Jian Pei and Wo-Shun Luk, "Email mining: Tasks, Common Techniques and tools", Knowledge and Information Systems, Issue 1, Vol. 41, Print ISSN:0219-1377, pp.1-31, 2014*
17. *Li, Wenjuan and Weizhi Meng. "An empirical study on email classification using supervised machine learning in real environments." IEEE International Conference, pp.7438-7443, IEEE, 2015*
18. *Alsmadi, Izzat and Ikdam Alhami. "Clustering and Classification of email contents", Journal of King Saud University-Computer and Information Sciences, Vol.27, Issue 1, doi:10.1016/j.jksuci.2014.03.014, pp. 46-57, 2015*